DOCUMENT RESUME

ED 170 379 TM 009 009

AUTHOR Stake, Robert E.
TITLE The Validity of District-Wide Test Means.
PUB DATE 10 Apr 79
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Rating; Achievement Tests; Communication Problems; *Curriculum Evaluation; Decision Making; Educational Diagnosis; Elementary Secondary Education; Group Tests; *Information Utilization; Performance Based Education; *School Districts; Science Curriculum; Scores; Test Interpretation; Test Results; *Test Validity

ABSTRACT

It is argued that achievement test means, based on group scores, or other district-wide summary scores do not provide valid measures for diagnosing curriculum weaknesses nor for initiating curriculum changes. Group scores are probably valid for some purposes, but should not be used as criterion variables for evaluating the effectiveness of instruction or for indicating the quality of education a school district provides. Based on 11 case studies of district science instruction in kindergarten through grade 12, it was found that a lack of communication existed between the district curriculum coordinators and testing personnel, that curriculum coordinators paid little attention to test results, and that testing people were not concerned about curriculum review. In another case study, researchers found no instances of curriculum change or recommendations for change based directly on student testing. Using group test means as a basis for curriculum diagnosis or program evaluation is a procedure which is not grounded in validity studies, but is based on experience and intuition. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Rev. 3/79

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert E. Stake

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE VALIDITY OF DISTRICT-WIDE TEST MEANS*

by

Robert E. Stake
University of Illinois at Urbana-Champaign

ED170379

In April of 1979, at this very minute, to the best of my knowledge, there is no established evidence of validity of achievement-test means or other district-wide summary scores for indicating what might be wrong with the curriculum, or for indicating how instruction might be changed to improve things. Most testing people know this and consider that only the formal evidence is lacking. They may be far too optimistic. The tests appear to have little diagnostic value for district decision-making even though claims are widely made that state testing programs and district assessment systems will provide a basis for improving the curriculum.

At this time our profession does not have the technical grounds to claim legitimately that there is diagnostic value in norm referenced testing or in criterion referenced testing -- on a school by school or district by district basis. When some of us looked for observational evidence inside the districts we found essentially none of the tests providing diagnostic information to curriculum coordinators -- nor even for that matter to teachers. But this paper is more about districts than teachers.

---

*Paper presented at Session 10A of the annual meeting of the National Council of Measurement in Education, San Francisco, April 10, 1979.

TM009 009

Achievement test information available to school district people is widely used to check whether scores are high enough -- as high as they have been, or as high here as they are elsewhere, or as high as common sense tells them scores should be. While discussing test means, curriculum coordinators may discover problems needing investigation or solutions needing trial. Many administrators, testing specialists and researchers consider such comparisons and reviews to be valuable. I think they are too. The test information itself can be useful not only politically for initiating funding changes, staffing changes, or changes in educational objectives, but the means sometimes do indicate actual change in the performance level of students.

In this presentation I am talking about group scores, not individual pupil scores. There is good evidence that for many purposes many achievement tests provide valid individual student information. The group scores are probably valid for some purposes as well, but at the time being, we do not know what the validities are. We cannot with confidence use these tests as criterion variables for evaluating the effectiveness of instruction or for indicating the quality of education a district is providing.

Many educators suspect this to be true, but as with the Emperor's new clothes ...

The confidence people have in using test-score means as an indicator at the school or district level is based on the professional experience of the educator using them. We do not have empirical studies that show that test means are effective indicators of school or district performance. It is not clear that such empirical studies are needed -- perhaps professional experience is the best basis for validating a performance indicator. However, people should not imply that the validity of these summary scores has been established by the test developers or by researchers, for it has not.

What I have said so far is that the present day achievement batteries do have an experiential or common-sense validity for indicating how certain district-wide performances are changing. They do not have an established validity for indicating what should be done about the changes. Some reviewers of district profiles look at the items missed most and suppose that that is where greater instructional stress should be placed. Maybe it should. There is no established validity, however, for the test score meanings implicit in that strategy, or any other that curriculum and testing people might now recommend.

Is it that the test means have no diagnostic validity or just that we have not studied them enough? I do not know. I suspect that they do not have such a diagnostic validity. Many educators have similar suspicions. The research of Hastings and Hotvedt* showed teachers spending little time going over the results of testing. If interested at all, the teachers looked to see whether scores were going up or down, and seldom looked further. Most teachers I talk to tell me they cannot expect to find information useful to them for instructing the children.

It is true that many teachers and many administrators do not understand the technical aspects of testing, but that probably is not the main reason they treat mean scores as containing little educational information. They know they are not going to find much that they did not already know about what the teaching ought to be.

* * * * *

---

*J. Thomas Hastings, Phillip J. Runkel and Dora E. Damrin. Effects on Use of Tests by Teachers Trained in a Summer Institute. Cooperative Research Project #702, Bureau of Educational Research, College of Education, University of Illinois at Urbana-Champaign, 1961.

Martyn D. Hotvedt. A Study of Standardized Test Use in Public Schools. Urbana: University of Illinois, 1974 (unpublished doctoral dissertation).

In the Fall of 1977 the nation received the report* of the study of College Board score decline. It was greatly revealing, one of the most meaningful educational research studies of recent years. We were told "no one cause could be found." That was the not-surprising but important insight of this study.

Here was a perfect opportunity for the testing researcher. The Scholastic Aptitude Test used by the College Board for applicants to college is perhaps the most carefully-built standardized test in the world. Its validity for predicting how well an examinee will perform later in college is well established. Between 1964 and 1976 national means (based annually on many thousands of examinees) had dropped slowly but steadily. Lots of research money was available for studying the drop. A distinguished social servant, Willard Wirtz, chaired a panel of outstanding educators and social scientists to do the work. They subcontracted a number of research papers. A staff associate, an old friend, Paul Barton, assured me it was a thoughtful, hard-working group.

Why were the mean scores going down? Were the schools doing a poorer job? Were more of the less-able high school students now seeking admission to college? Were students no longer trying their best to get high test scores? How valid was the drop in test scores for evaluating American high schools?

The panel established that the tests were suitably equivalent from year to year. The drop in scores was real, but what did it mean? They found the population of student examinees had changed, especially in the 1960's, but this was not seen to be the outstanding cause of the score decline. In fact, no outstanding cause was found, just numerous contributing factors. Yes, the

---

*Willard Wirtz et al. On Further Examination. Report to the Advisory Panel on the Scholastic Aptitude Test Score Decline. New York: College Entrance Examination Board, 1977.

attitude toward testing had changed. Yes, the attitudes toward school had changed. Yes, curriculum reform had changed the aim of teaching to some degree away from what the tests emphasized. Etc., etc. ... What was clear was that this great aggregation of student performance data, analyzed at length by the most able of scholars, yielded no answers to evaluation questions about the quality of teaching in the American schools over those years.

The College Board situation centered on the analysis of a scholastic aptitude test rather than an achievement test, but I believe the results would have been no different had they studied annual scores from any standardized achievement test in any particular place. If that is the case, is it reasonable to suppose that a single testing consultant, or district research team, or committee of teachers can find causes for student performance by looking at district-wide test scores? It is my belief that changes in standardized test-score means are determined by causes too complex and too remote for present-day analysis. Each mean is a temporary end-point for a myriad of potential "causes." Seldom, I think, will those means be suitable criterion scores for evaluative research studies.

* * * * *

For two years a team of us based at the University of Illinois were studying science teaching in grades K-12 in the United States.* Our method primarily was naturalistic, yielding eleven very nice case studies of district science instruction. We became interested in the role of the district curriculum-coordinator, though our attention was not usually on the district office. It was apparent to us that there was not much communication between

---

*Robert E. Stake and John A. Easley, Jr. Case Studies in Science Education. Washington D. C.: U. S. Government Printing Office, 1978 (2 volumes + executive summary).

the curriculum people and the testing people -- no jurisdictional disputes or anything like that, just a lack of common territory. The curriculum coordinators we encountered did not pay much attention to test results other than to check on level of performance, overall or special subscores. The testing people sometimes were involved in microanalysis of specific lesson objectives, but almost never were concerned about major review of the curriculum.*

Sometimes the curriculum coordinators confided to us that they had strong reservations about an objectives-based curriculum and competency-based testing, but had to go along to protect "their turf, their subject matter" when funding or staffing decisions were being made. It was apparent to us that many administrators presumed the validity of the testing program and even subtly encouraged the testing people to over-promise what the tests could do to evaluate the curriculum.

To study the situation further I persuaded a mature and competent graduate assistant, Mr. Norman Bowman, to find a good example of a school district where testing information was being used for curriculum evaluation and development. He first looked for consultants or trouble shooters, people who had reputations for discerning, from aggregated test data, a suitable basis for modifying instruction. Some administrators told him no such people existed; some said curriculum analysis based on testing was accomplished by committee, not by an individual; and some said it was a common ability among master teachers. When he interviewed some of the nominees he found only

---

*In her study of district Research and Evaluation Offices, Catherine Lyon of the UCLA Center for the Study of Evaluation found testing people more linked to administrative offices than instructional offices, and consequently less involved in curricular questions.

people who used tests as an indicator of rise or fall in performance, not one who used tests for group diagnostic purposes.

He then located himself in one suburban district where the schools were well funded, the testing program active, where a nationally recognized consultant in testing was regularly involved, and where the superintendent was both widely respected and committed to using technical resources (such as testing) to benefit instruction. Mr. Bowman observed, interviewed and studied the records for six weeks. He did not find a single instance of curricular change (or recommendation for change) directly based upon student testing. Test performances sometimes signalled it was time for change but not what changes should be made. Bowman's search of course was a small one, and he could have been somehow blinded to what was there. But it seems safer to me to conclude that curriculum change based on test scores is a very rare event.*

Across the country today there are many state and district assessment programs based on the assumption that district personnel are ready and able to use group scores for diagnostic purposes. It should surprise some people that Mr. Bowman was not able to document how even a few such people are presently using test scores to direct the improvement of curricula.

Mr. Bowman was aware that some districts are developing new curricular materials to be coordinated with test questions that will be asked in the future. Numerous large-scale curriculum projects have been so organized: Bill Gorth's Project CAM and Jerrold Zacharias' Project TORQUE come to mind.**

---

*Norman F. Bowman. A Search for Instances of District Use of Aggregated Test Data for Curriculum Improvement. Urbana: University of Illinois, 1979 (unpublished doctoral dissertation).

**Bill Gorth is at the University of Massachusetts. Jerrold Zacharias is at the Educational Development Corporation of Newton, Massachusetts.

But this is the development of curricula based on notions of what the students should be able to answer, not on what the students did answer. What those folks have done has little to do with building curricula from district means actually obtained.

It is difficult to estimate the validity of district means for a purpose that is not being pursued. But the question is of interest because enormous investments are being made in testing programs partly justified by claims of diagnostic validity.

* * * * *

Perhaps we should consider for a moment whether or not a mean could represent the quality of district instruction. Can we think of examples when the mean poorly indicates the quality of the whole?

> Is the mean strength of earth-filled dams an indication of the quality of work of the Army Corps of Engineers?
>
> Is the mean ripeness of tomatoes at the supermarket a good indication of their quality?
>
> Is the mean competence of program evaluators a good indication of the quality of program evaluation?

It is obvious that means can give too little information. Even though the arithmetic mean is statistically sensitive to each individual case, it may not be sensitive enough for our purposes. Means also draw our attention more to the entirety of the group rather than to the circumstance of each and all individual members within the group. Should any one dam break, should the tomato any diner gets be rotten, should any one program evaluator be without an understanding of educational process, the whole is greatly affected. Education is very much a business of attending to the individual. Excess in one regard is not nullified by excess of the opposite kind. Efforts to evaluate education by concentration on means not only are simplistic, but may misdirect our

attention as well. So we might have reservations about ever validating mean test scores as indicators of the quality of district, or school, or even classroom instruction.

* * * * *

Lee Cronbach has contributed as much as any person to the clarification of validity of educational measurements. He has stressed that validity lies in the meaning we attach to the scores.* A given test will have, not a single validity, but many validities, depending on the uses to be made of the scores, depending on the student populations tested, depending on which test information is being used. It follows that the mean score for a district does not necessarily have the same validities that an individual test score has.

Many districts use an achievement test such as the Comprehensive Test of Basic Skills. It has an established validity for indicating certain achievement of certain students. In the experience of many educators the individual CTBS score is a good indicator of individual student attainment. It has occasionally been statistically validated by researchers. Never, to my knowledge, has the validity of the mean of the test been established empirically -- that is, to have researchers substantiate with data that the mean indicates something about the quality of teaching or the quality of learning in that district.

The validity of a mean cannot be calculated from the validity of individual scores -- even though the error of measurement of a mean can be calculated from the error of measurement of single scores. Validity is based on actual use. One cannot suppose that the meanings of the district score will be the same as those attached to the individual student score.

---

*Lee J. Cronbach. Test Validation. In R. L. Thorndike (Ed.) Educational Measurement, second edition. Washington D. C.: American Council on Education, 1971.

The greatest apprehension about test means probably occurs when they are published in the newspapers. The meaning the public often gives to a change in scores is that it indicates a general change in quality of student performance. A great number of changes occur because the group of youngsters tested changed composition or because the test changed somehow. Changes in student attainment are sometimes reflected on the test too, of course.

By and large, standardized test items draw out some of the competence the student has acquired, but only a small part. If the teachers have been preparing the children directly for the test then the performance should reflect that learning pretty well -- but it will not show how well they are acquiring all the basic skills, the general knowledge, and particularly the complex associations we expect of human beings, young and old.

If the district's teachers have been doing what teachers traditionally have done to teach skills, knowledge and experiential associations, then the tests will test only a small portion. Most testing people believe that these limited tests still are good indicators of the general achievement of school work. They point to the high correlation between tests covering different knowledges and competencies, but they gloss over the fact that the correlations tell us whether or not the same people do best, not whether anyone does as well as they ought to. The fact of the matter is that the district achievement test mean may be a very poor indicator of how well the curriculum has been chosen, how well it has been taught and how well it has been learned.

We just don't know. The validity of these means has not been established. They mean something of course, but what? A superintendent or curriculum coordinator who changes the curriculum because of student performance means is either very experienced in matters of curriculum and instruction, or a good politician, or a gambler, and possibly all three.

Test means are products of a highly developed technology, but using them for curriculum diagnosis or program evaluation is an act of faith, a reliance on experience and intuition, not grounded in validity studies that researchers have completed.